HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2015 November 11.

Published in final edited form as:

Stat Med. 2013 March 15; 32(6): 1016–1026. doi:10.1002/sim.5575.

A Semiparametric Recurrent Events Model with Time-varying Coefficients

Zhangsheng Yu[1,*], **Lei Liu**[2], **Dawn M. Bravata**[3,4], **Linda S. Williams**[3,4], and **Robert S. Tepper**[5]

[1]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, U.S.A.

[2]Department of Preventive Medicine, Northwestern University, Chicago, IL

[3]Richard L. Roudebush VA Medical Center, Indianapolis, IN

[4]Regenstrief Institute, Inc., Indianapolis, IN

[5]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN

SUMMARY

We consider a recurrent events model with time-varying coefficients motivated by two clinical applications. A random effects (Gaussian frailty) model is used to describe the intensity of recurrent events. The model can accommodate both time-varying and time-constant coefficients. The penalized spline method is used to estimate the time-varying coefficients. Laplace approximation is used to evaluate the penalized likelihood without a closed form. The smoothing parameters are estimated in a similar way to variance components. We conduct simulations to evaluate the performance of the estimates for both time-varying and time-independent coefficients. We apply this method to analyze two data sets: a stroke study and a child wheeze study.

Keywords

Penalized spline; variance components; survival analysis; semiparametric regression

1. Introduction

In longitudinal studies, events of interest are often observed for multiple times, for example, hospital readmissions after stroke, or episodes of wheezing among young children. This type of recurrent events data have been studied extensively using marginal (rate or mean) models [1]) and random effects (frailty) models [2]. In these models, it is often assumed that the risk factor's effect is unchanged over the follow-up period. This convenient assumption may not be realistic, and may lead to bias in the coefficient estimates.

Varying-coefficient models for survival analysis have been studied by many authors. Cai et al. used a local polynomial method for a *marginal* survival model with time-varying

*Correspondence to: yuz@iupui.edu.

coefficients [3]. Yu and Lin studied a semiparametric *marginal* model with time-varying coefficients for clustered survival data [4]. Sun and colleagues studied a recurrent events model with time-varying coefficients using a marginal modeling approach [5, 6]. Other related work includes [7, 8, 9, 10, 11]. But to the best of our knowledge, no work has been published for a recurrent events model with time-varying coefficients under the random effects (*frailty*) model framework.

Our work was motivated by two clinical studies. A childhood wheeze study was designed to examine the effect of airway reactivity during infancy on recurrent wheezing. For many children, more than one episode of wheezing were observed during the follow-up. To accommodate the correlation among the occurrences of multiple episodes, a random effects (frailty) model can be used. While the wheezing symptom at an older age is most often related to airway reactivity, this is not the case in very early life. Clinical experience indicates that wheezing in very early life is caused by multiple factors that may not be related to airway reactivity. Therefore, it is important to accommodate the age- or time-varying effect of an infant's airway reactivity, which motivated us to develop a recurrent events frailty model with time-varying coefficients. The proposed model demonstrates that airway reactivity measured at baseline has a significant effect on wheezing at a later age (> 52 months), but not at an early age (< 52 months). In another study, we are interested in the effect of stroke care quality on readmission rate for stroke patients. We suspect that the risk of readmissions in a short term after discharge is mainly determined by the severeness of the disease, while a better stroke care may reduce the long term risk. Therefore, we apply the proposed models to analyze the readmission rate with time-varying coefficients for stroke care quality.

In our model, a Gaussian frailty is used to characterize the correlation among recurrent events. To approximate the marginal likelihood with an integral due to the unobserved random effects, we employ the Laplace approximation in a similar fashion to generalized linear mixed models [12]. We use the penalized spline method to characterize the time-varying coefficient function. The variance components and smoothing parameters are estimated using a likelihood method.

The remainder of the article is arranged as follows. We present the model in Section 2, and propose the estimation method in Section 3. The estimation method is evaluated by a simulation study in Section 4. In Section 5, we apply our model to data from two prospective cohort studies. We conclude with a discussion in Section 6.

2. Models

For the ith subject ($i = 1, 2, \ldots, n$), we observe r_i recurrent events times $0 < R_{i1} < R_{i2} < \cdots < R_{ir_i}$ before being censored at time C_i, which is independent of R_{ij}. Let $\boldsymbol{Z_i}$ be a $p-$dimensional covariate vector with time-varying coefficients, and $\boldsymbol{X_i}$ be a $q-$dimensional covariate vector with constant coefficients. Denote the observed recurrent events process as $N_i^R(t)=\sum_{j=1}^{r_i} I\{R_{ij} \le t\}$, at risk process as $Y_i(t) = I\{C_i \ge t\}$. The observed information of

the *i*th subject at time *t* is $\mathbf{O}_i(t)=\{Y_i(u), N_i^R(u), \boldsymbol{Z_i}, \boldsymbol{X_i}, 0 \le u \le t\}$. The filtration generated by the observed information is $\mathscr{F}(t) = \sigma\{\mathbf{O}_i(s), 0 \le s \le t, i = 1, \cdots, n\}$.

We now formulate the intensity function for recurrent events as

$$P(dN_i^R(t)=1|\mathscr{F}_{t^-})=Y_i(t)dR_i(t),$$

with

$$dR_i(t)=\exp\{\boldsymbol{Z_i^T}\boldsymbol{\beta}(t)+\boldsymbol{X_i^T}\boldsymbol{\alpha}+\nu_i\}r_0(t)dt, \quad (1)$$

where $r_0(t)$ is the baseline intensity of recurrent events, $\beta(t) = (\beta_1(t), \cdots, \beta_p(t))^T$, and we use $R_0(t)$ to denote the cumulative baseline intensity function. The unobserved random effect (or frailty) is denoted by ν_i, which is used to capture the correlation among recurrent events. Denote the density of ν_is by $\varphi_\sigma(\cdot)$. In this paper, we assume ν_is to be independently Gaussian distributed with mean 0 and variance σ^2. We use $\beta(t)$ to denote the unspecified smooth time-varying coefficients, whose functional form is of interest. Using these notations, we write the complete likelihood of $\{(\mathbf{O}_i,\nu_i), i = 1, \cdots, n\}$ as

$$\log\prod_{i=1}^n L(\mathbf{O}_i, \nu_i)=\log\prod_{i=1}^n L(\mathbf{O}_i|\nu_i)L(\nu_i)=\sum_{i=1}^n l_i+\sum_{i=1}^n \log\phi_\sigma(\nu_i),$$

where

$$l_i=\left\{\sum_{j=1}^{r_i}[\nu_i+\boldsymbol{Z_i^T}\boldsymbol{\beta}(R_{ij})+\boldsymbol{X_i^T}\boldsymbol{\alpha}+\log r_0(R_{ij})] - \int_0^\infty Y_i(t)\exp\{\boldsymbol{Z_i^T}\boldsymbol{\beta}(t)+\boldsymbol{X_i^T}\boldsymbol{\alpha}+\nu_i\}r_0(t)dt\right\}.$$

The complete likelihood involves ν_i, and the time-varying coefficients $\beta(t)$ with an infinite number of parameters. We will employ the penalized spline method to describe the time-varying coefficients $\beta(t)$. Laplace approximation will be adopted for evaluating the complete likelihood. The estimation procedure is given in next section, which can be easily extended to accommodate time-dependent covariates ***X*(*t*)** and ***Z*(*t*)**.

3. Estimation method

3.1. Penalized Spline

We use a penalized spline to estimate the time-varying coefficients. The penalized spline method uses a penalty term to control the smoothness of the spline estimate but includes a smaller number of knots to reduce the computational load, compared to the smoothing spline method. Specifically, we model the nonparametric function $\beta_l(t)$ through a cubic spline with basis functions $\{B_1(t), B_2(t), \cdots, B_M(t)\}$, where M is the number of spline basis functions. The number and the shape of the basis functions are determined by the number and location of knots. With a penalized spline, one can choose a larger number of knots without introducing much more variation. But there should be little advantage to use more than 10–

20 knots as Gray recommended [13]. We will use 8 knots in our simulation study for demonstration. It is suggested that the knots are placed so that an equal number of events happen within each interval.

We write $\beta_l(t)=\sum_{m=1}^{M}\eta_{l,m}B_m(t)=\boldsymbol{\eta_l^T B}(t)$ for $l = 1, \cdots, p$, where $\boldsymbol{\eta_l} = (\eta_{l,1}, \cdots, \eta_{l,M})^T$ and $\boldsymbol{B}(t) = (B_1(t), \cdots, B_M(t))^T$. The penalized log-likelihood with a penalized spline for $\{(\mathbf{O}_i,\nu_i), i = 1, \cdots, n\}$ is

$$pl=\sum_{i=1}^{n}l_i+\sum_{i=1}^{n}\log\phi_\sigma(\nu_i)-\sum_{l=1}^{p}\frac{1}{2\rho_l}\boldsymbol{\eta_l^T P\eta_l}$$

where $\mathbf{P} = \int \boldsymbol{B}^{(2)}(s)\boldsymbol{B}^{(2)}(s)^T\,ds$, the vector $\boldsymbol{B}^{(2)}(s)=\{\boldsymbol{B}_1^{(2)}(s),\cdots,\boldsymbol{B}_M^{(2)}(s)\}^T$ is the second derivatives of the B-spline basis, and ρ_l is the smoothing parameter. Note that the roughness of the likelihood function was penalized by subtracting the integral of the squared second derivative function. In practice, the cubic spline basis functions and the penalty matrix can be easily generated by the *fda* package in R-project or matlab.

3.2. Evaluating the Marginal Likelihood

To estimate the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\eta} = (\boldsymbol{\eta_1}, \cdots, \boldsymbol{\eta_p})$ in the penalized likelihood, the frailty has to be integrated out. The logarithm of the penalized marginal likelihood is

$$pml=-\sum_{l=1}^{p}\frac{1}{2\rho_l}\boldsymbol{\eta_l^T P\eta_l}+\log\left[\int_{-\infty}^{\infty}\prod_{i=1}^{n}\exp\{l_i+\log\phi_\sigma(\nu_i)\}d\boldsymbol{\nu}\right] \quad (2)$$

where $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_n)^T$. There is no closed form for the logarithm integral part in (2). To evaluate (2), we use the Laplace approximation for integral calculation by following Breslow and Clayton's derivation for generalized linear mixed models [12], i.e.,

$$\log\{c|\boldsymbol{D}|^{-1/2}\int e^{-K(\boldsymbol{\nu})}d\boldsymbol{\nu}\}\approx-\tfrac{1}{2}\log|\boldsymbol{D}|-\tfrac{1}{2}\log|K''(\tilde{\boldsymbol{\nu}})|-K(\tilde{\boldsymbol{\nu}})+\text{constant}, \quad (3)$$

where $\boldsymbol{D}$ is the covariance matrix of $\boldsymbol{\nu}$, $\tilde{\boldsymbol{\nu}}$ is the solution to the first derivative $K'(\boldsymbol{\nu}) = 0$, and $K''(\cdot)$ is the second derivative of $K(\cdot)$ with respect to $\boldsymbol{\nu}$. After applying the approximation on the logarithm part of *pml*, we have

$$K(\boldsymbol{\nu})=-\sum_{i=1}^{n}l_i+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\nu_i^2$$

by noting that $\varphi_\sigma(\nu_i)$ is normal distributed with mean 0 and variance σ^2.

Assuming that the first two terms in (3) vary little when $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ changes, as indicated by Breslow and Clayton and Ripatti and Palgram [12, 14], the logarithm part of (2) can be further approximated by $-K(\tilde{\boldsymbol{\nu}})$. Thus, we have

$$pml \approx \sum_{i=1}^{n} l_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\nu_i^2 - \sum_{l=1}^{p}\frac{1}{2\rho_l}\boldsymbol{\eta}_l^T\boldsymbol{P}\boldsymbol{\eta}_l$$

To profile out the the cumulative baseline intensity $R_0(t)$, we take the derivative of above *pml* with respect to $r_0(R_{ij})$, and obtain

$$\hat{r}_0(R_{ij}) = \frac{1}{\sum_k I\{C_k \geq R_{ij}\} e^{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(\boldsymbol{R_{ij}})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k}} \quad (4)$$

Substituting the solution back into the penalized marginal likelihood, we simplify the *pml* to

$$\begin{aligned} pml = & \sum_{i=1}^{n}\sum_{j=1}^{r_i}\{\nu_i \\ & + \boldsymbol{Z_i^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij}) + \boldsymbol{X_i^T}\boldsymbol{\alpha} - \log\sum_k I\{R_{ij} \leq C_k\} e^{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(\boldsymbol{R_{ij}})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k}\Big\} + \sum_{i=1}^{n}\log\phi_\sigma(\nu_i) \\ & - \sum_{l=1}^{p}\frac{1}{2\rho_l}\boldsymbol{\eta}_l^T\boldsymbol{P}_{\boldsymbol{\eta}_l} \\ = & \sum_{i=1}^{n} l_i + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\nu_i^2 \\ & - \sum_{l=1}^{p}\frac{1}{2\rho_l}\boldsymbol{\eta}_l^T\boldsymbol{P}_{\boldsymbol{\eta}_l}, \end{aligned} \quad (5)$$

Note that $\boldsymbol{Z}_i^T\boldsymbol{\eta}^T\boldsymbol{B}(R_{ls}) = \sum_{l=1}^{p} Z_{il}\boldsymbol{\eta}_l^T\boldsymbol{B}(R_{ls})$.

We then derive the estimating equations of $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ by taking the derivative of (5) with respected to $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ as follows:

$$S_{\boldsymbol{\eta_s}} = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\{Z_{is}\boldsymbol{B}(R_{ij}) - \frac{\sum_k I\{C_k \geq R_{ij}\} Z_{ks}\boldsymbol{B}(R_{ij}) exp\{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k\}}{\sum_k I\{C_k \geq R_{ij}\} exp\{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k\}}\} - \frac{1}{\rho_s}\boldsymbol{P}\boldsymbol{\eta_s^T}, \quad (6)$$

for $s = 1, \cdots, p$;

$$S_{\boldsymbol{\alpha}} = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\{\boldsymbol{X_i^T} - \frac{\sum_k I\{C_k \geq R_{ij}\}\boldsymbol{X_k^T} exp\{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k\}}{\sum_k I\{C_k \geq R_{ij}\} exp\{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k\}}\} \quad (7)$$

and

$$S_\nu = (r_1, \cdots, r_n)^T - \sum_{i=1}^{n}\sum_{j=1}^{r_i}\frac{\{I_{ij1}, \cdots, I_{ijn}\}^T}{\sum_k I_{ijk}} - \frac{\boldsymbol{\nu}}{\sigma^2}, \quad (8)$$

where $I_{ijk} = I\{C_k \geq R_{ij}\} e^{\boldsymbol{Z_k^T}\boldsymbol{\eta}^T\boldsymbol{B}(R_{ij})+\boldsymbol{X_k^T}\boldsymbol{\alpha}+\nu_k}$.

The estimates of regression coefficients can be obtained by solving the estimating equations (6, 7, 8) using the Newton-Raphson algorithm. After obtaining the estimates of $\boldsymbol{\eta}$, the varying-coefficient estimates are then $\hat{\beta}_l(t)=\sum_{m=1}^{M}\hat{\eta}_{lm}B_m(t)$. We propose using $-I(\boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\nu})$, the inverse of the negative second derivative of *pml* with respect to $(\boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\nu})$, as their covariance estimate. The variance of $\hat{\beta}_l(t)$ can then be estimated by $\widehat{Var}\{\hat{\beta}_l(t)\} = \boldsymbol{B}(t)^T\boldsymbol{V}_{\boldsymbol{\eta}\boldsymbol{\eta}_l}\boldsymbol{B}(t)$, where $\boldsymbol{V}_{\boldsymbol{\eta}\boldsymbol{\eta}_l}$ is the submatrix of $-I(\boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\nu})$ corresponding to $\boldsymbol{\eta}_l$. Estimation for $\boldsymbol{\alpha}$ and its variance can be done similarly. The baseline intensity functions $r_0(t)$ can be estimated by plugging $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\nu}}_i)$ back into (4).

3.3. Inferences on Variance Components and Smoothing Parameters

Estimation of variance components σ using a restricted maximum likelihood method (REML) has been proposed by Ripatti and Palgrem [14]. For a shared frailty model where the covariance of $\boldsymbol{\nu}$ is a diagonal matrix with σ^2 on the diagonal, the variance component can be estimated as

$$\hat{\sigma}^2=\frac{\hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}+tr\{[K''(\hat{\boldsymbol{\nu}})^{-1}]\}}{n_c}, \quad (9)$$

where n_c is the number of clusters (or subjects), and $tr(\cdot)$ is the trace operation of a matrix. The information matrix $K''(\boldsymbol{\nu})$ requires the estimate of the baseline intensity function.

In general, estimation of the smoothing parameter ρ_l can be performed using the cross-validation method [15]. In survival analysis, there is not a well accepted cross-validation score, to our best knowledge. Here, we suggest estimating ρ_l in a similar way to variance components [16]. Equation (9) includes the number of parameters (random effects) in the denominator, while the numerator include two terms: sum of square of estimators and trace of covariance matrix. Therefore, we mimic the estimating equations for variance component σ^2 in (9) as follows:

$$\hat{\rho}_l=\frac{\sum_{i=1}^{n}\sum_{j=1}^{r_i}\hat{\beta}_l(R_{ij})^2+tr\{Cov(\hat{\beta}_l(R_{ij}))\}}{M},$$

for $l = 1, \cdots, p$, where $Cov(\hat{\beta}(R_{ij}))$ is the covariance matrix of the nonparametric function, M is the number of basis functions for the penalized spline. The performance of the resulting estimates will be evaluated in the next section.

4. Simulation

In this section, we used simulation to evaluate the performance of the proposed estimates. We considered an intensity model for recurrent events as follows:

$$dR_i(t)=exp\{Z_i\beta(t)+X_{1i}\alpha_1+X_{2i}\alpha_2+\nu_i\}r_0(t)dt,$$

where ν_i was normal distributed with mean 0 and variance 0.5. The covariates Z_i and X_{i1} were generated as independent uniformly distributed random variables on [0, 1]. X_{i2} took

value 0 or 1 with an equal probability. We adopted a baseline intensity function $r_0(t) = \sqrt{t}$ to simulate an increasing risk. The censoring times were generated as an exponential distributed variable independent of the recurrent events process. Event processes were censored at the maximum follow-up time, 4, if not censored before.

We simulated intensity models in two parameter settings. In the first setting, we generated the time-varying coefficient function as $\beta(t) = (sin(t \times 3\pi/8) + 1)/4$ as showed in panel (a) of Figure 1, the function increased first and then diminished to 0 along the time. We set $(\alpha_1, \alpha_2) = (1, 1)$ in Setting I. To examine the performance of the proposed estimate when time-varying coefficient function has a stronger curvature, we ran simulation studies with $\beta(t) = (2 \times beta(t/4, 8, 8) + beta(t/4, 5, 5))/9$ in Setting II, where *beta* represented the density function of a beta distribution. The shape of the true function was shown in Figure 2. We setted $(\alpha_1, \alpha_2) = (2, 2)$ in Setting II. In each setting, we also evaluated the performance when the numbers of subjects were 200 and 100, and the numbers of event per subjects were 5 and 3, respectively. Under each simulation, we generated 800 data sets for evaluation. We implemented the estimation procedure in R-project. The code is available upon request.

Figure 1 shows the simulation results of the time-varying coefficients in Setting I with 200 subjects and 5 events per subject on average. In the left panel, the solid line is the true function and the dotted line is the mean of the estimated function over 800 replicates. The bias of the estimated function is very small. The right panel shows that the pointwise empirical coverage probabilities, calculated at each of 100 equally spaced grid points. The coverage probabilities are around 95% for most of the follow-up period. The average coverage probability over 100 grid points in the follow-up period is 96.4%. To test how sensitive the estimate to the number of events per subject, we also run the simulation with a sample size of 200 but 3 events per subject. The time-varying coefficient estimate shows very little bias (not shown) and the empirical coverage probability is 96.5% on average. Meanwhile, we also run simulations with smaller sample size of 100 and 3 and 5 events per subject. The varying-coefficient estimate shows a similar or slightly larger bias (not shown).

Figure 2 shows the estimate of the time-varying coefficient function in Setting II. In the left panel, the dotted line is the mean of estimated function and the solid line is the true function. The bias is generally small over the whole range of time. The right panel shows that the empirical coverage probabilities are approximately 95% with an average coverage probability of 96.5%. Simulation with 3 events per subjects displays a similar performance and study with 100 subjects shows a slightly larger bias. These results demonstrate that the time-varying coefficient estimates have little bias and the pointwise coverage probabilities are close to the nominal level for a moderate sample size 200 and a small number of events per subjects of 3. In the settings with small sample size of 100, bias is similar or slightly larger but the estimate still catch the shape of the true function well.

Table I lists the simulation results for the parametric coefficients and variance components. In Setting I, with 200 subjects and 5 events per subject, the estimates for (α_1, α_2) have about 3% biases. The empirical coverage probabilities of 95% confidence intervals using the estimated standard error are 94.4% and 94.0%, which are close to the nominal level. The true value of the variance component σ^2 is 0.5 and the mean estimated value is 0.465 which

is biased slightly downward. This bias is similar to Ripatti and Palgrem where they used Laplace approximation for a frailty model with parametric coefficients only [14]. When the number of events per subjects is 3 and the sample size is 100, the coefficient estimates show similar bias, but with a larger standard error due to the smaller sample size. The coverage probabilities of the 95% confidence interval for these parameters are also close to the nominal level. In Setting II, the estimates of parametric coefficients show similar performance.

In summary, the bias of the proposed estimates for the time-varying function and parametric coefficients are small even with a small sample size of 100 or a number of events per subject of 3. The confidence intervals based on the proposed standard error estimates have appropriate coverage probabilities. The variance component estimate has a modest downward bias and should be used with caution.

5. Application

5.1. A Child Wheezing Study

Older children with asthma have frequent episodes of wheezing and exhibit airway hyper-reactivity when assessed by bronchial challenge testing with inhaled methacholine. Airway reactivity can be characterized by log PC_{30}, the logarithm of the provocative concentration of methacholine required to decrease airway function by 30%. A lower value indicates greater airway reactivity. Wheeze is also a common respiratory symptom for infants and toddlers; however, most of these subjects would have episodes of wheezing in very early life, and not develop childhood asthma. We evaluated whether measurement of airway reactivity very early in life could predict the risk of wheezing at an older age, which would likely represent asthma, rather than transient wheezing of infancy. In a cohort of 116 infants at high risk for developing childhood asthma, PC_{30} was measured upon enrollment, prior to any episodes of wheezing [17]. The subsequent episodes of wheezing were assessed by monthly telephone contacts with families. Eighty-nine subjects completed the assessment of airway reactivity upon entry to the study and were included in the analysis. Among them, 52% were boys, and 11.2% had mothers smoking during pregnancy. The mean value for log PC_{30} was 0.809 with standard deviation of 1.0017.

Among the 89 subjects (with median follow-up time 64.8 months), a total of 663 wheezing events were recorded. To accommodate correlation among the multiple events, we applied the recurrent events model with a frailty term to analyze the effect of airway reactivity upon entry on the recurrent wheezing risk. We first fitted a recurrent events model with constant effect for log PC_{30}, gender, and mother smoking during pregnancy. A higher log PC_{30}, or lower airway reactivity, was associated with a lower risk of wheezing (hazard ratio = 0.899), but the relation was not statistically significant (p-value = 0.530). This is consistent with clinical experience that early wheezing is a multiple-cause symptom, which may not be related to airway reactivity. However, the wheezing symptom at an older age is more related to childhood asthma. Therefore, the effect of airway reactivity log PC_{30} at baseline may be age-varying. Treating the initial airway reactivity effect as a constant over time would bias the estimate. We then applied our model with time-varying coefficients for log(PC_{30}) effect. Figure 3 illustrates the effect of log PC_{30} for both the time-varying coefficient model (solid

line) and the constant coefficient model (dashed line). At the very early age (younger than 12 months), the curvature and wide confidence interval is due to the smaller number of wheezing events. In general, the baseline log PC_{30} value had little effect on the risk of wheezing before 50 months of age and started to have significant influence at around age 52 months. Subjects with a higher log PC_{30} (i.e. lower airway reactivity) had a lower risk of recurrent wheezing after age 52 months and the relative risk decreased with age afterwards. This finding suggested that the greater airway reactivity at baseline was a strong indicator of persistent wheezers (those who had recurrent episodes after around 4 years of age) or later-onset wheezers (those who started wheezing after 4 years old). We also adjusted for gender and maternal smoking during pregnancy as risk factors for wheezing (see Table 2). Children with mother pregnancy smoking had a 82% higher risk than those without (p-value<0.0001). Boys had higher risk than girls. The estimate of σ is 0.12. Overall, we demonstrated that airway hyper-reactivity prior to episodes of wheezing was a significant risk factor for wheezing at a later age.

5.2. A Stroke Study

Stroke affects 795,000 people each year in the United States [18]. Within the veteran affairs (VA) health system, approximately 6,000 veterans were admitted to a VA facility for acute ischemic stroke in fiscal year 2007 [19]. A natural sample of about 5, 000 veterans hospitalized in a VA medical center with acute ischemic stroke were evaluated with a chart review. The quality of inpatient stroke care was assessed using 14 indicators, which included those reflecting early hospital care period: dysphagia screening before oral intake (documentation of stroke severity using the NIH Stroke Scale, and thrombolysis administration), in-hospital care (antithrombotic therapy by the end of hospital day two, deep vein thrombosis prophylaxis, early ambulation, fall risk assessment, pressure ulcer risk assessment, and rehabilitation consultation based on the Functional Improvement Measurement documentation), and care at discharge (antithrombotic therapy at discharge, atrial fibrillation management, lipid management, stroke education, and smoking cessation counseling). A composite performance score of care quality received by each patients was calculated as the number of processes performed over the number processes for which the patient was eligible.

Among the veterans admitted to VA facilities in 2007, about 1300 were excluded since they were not admitted for an acute stroke; 3768 patients remained in the cohort. Patients who died while in hospital, were discharged to hospice, or received comfort care only, were excluded due to the potential very short follow-up time after discharge. The readmission records of all subjects were obtained through the VA administrative data. Among 3730 patients included in the final analysis, 2127 patients had at least one readmission to a VA hospital with a total number of 5168 readmissions (median: 2; range 1 – 26). The mean age at baseline was 67 years. There were 2.5% female patients, 67% Caucasians and 25% African Americans. The mean Charlson comorbidity score was 4.9 for patients with readmission and 4.4 for patient without readmission. The median follow-up time was 782 days, while 1031 died during the follow-up period.

We are interested in how the composite stroke performance score and comorbidity affect the readmission rate. To accommodate the multiple readmission events from the same patients, we first fit a naive frailty model with constant coefficients for performance score, Charlson's score, and smoking status. A higher performance score was associated with a lower risk of readmission (hazard ratio = exp(−0.957) = 0.384 with p-value = 0.009). We suspect that the risk of readmission in a short term after discharge is mainly determined by the severeness of the disease, but not by the stroke care quality. Therefore, we further analyzed the readmission by including the stroke care performance score with a time-varying coefficient and adjusting for Charlson's score and smoking status. The coefficient functions from both models are shown in Figure 4. Overall, the higher performance score corresponded to a lower risk (negative coefficient) of readmission, but it was not significantly in the short term according to the time-varying coefficient estimate. The effect increased over time after discharge. Starting at 300 days post discharge, the performance score showed a significant effect on the risk of readmission. This finding indicated that increasing in-hospital quality of care had a significant beneficial effect on the post discharge outcome of readmission in the long term. Compared with the naive analysis, the time-varying coefficient analysis provided a more detailed interpretation of the performance score effect over time. The estimates of Charlson's score, smoking status effect and σ are summarized in Table 3. A higher Charlson's score corresponded to a higher risk (Hazard ratio 1.208, p-value < 0.0010). Smokers had a higher risk, but not significantly.

6. Discussion

We have proposed a new model for the recurrent events data with both time-varying and constant coefficients. The penalized spline method was used to estimate the time-varying coefficient. We selected smoothing parameters by treating them as variance components. Simulation demonstrated a good performance of the proposed estimation method. We applied the method to two data sets and found significant time-varying coefficients in both cases. This demonstrated the wide application of our method in clinical studies.

The finding of the time-varying effect of airway reactivity on wheezing is consistent with the rapid childhood lung development. This renders the importance of time-varying coefficient models in the analysis of childhood asthma study. The finding that the quality of inpatient stroke care is related to late but not early reduction in readmission may reflect risk factor modification processes that begin while the patient is hospitalized (e.g. starting a cholesterol lowering medication) but take a long time to affect the outcome of interest (readmission). This is important since at present most hospital-level quality indicators, including those reported publically by the federal Centers for Medicare and Medicaid Services (available at www.hospitalcompare.hhs.gov), are evaluating the association with events in a shorter time horizon, typically 30-day mortality or readmission rates.

The proposed approach analyzed the time from beginning of the follow-up period to each of recurrent events. Meanwhile, gaps or waiting times also have been used for recurrent event modeling [20, 21], which is desirable when an individual is cured or a system is repaired to a similar state after each event. It is of future interest to study the time-varying coefficient estimation for gap time modelling.

In this paper we used a Gaussian frailty to model the correlation. As suggested by a reviewer, we also evaluated our estimates in the settings when the random effect was generated from other distribution (e.g., Gamma), but was assumed to be log normal in the simulation study, along the same lines of Huang and Liu [21].We found that our method still performed reasonably well under the misspecification of the frailty distribution assumption with Gamma (not shown). This is consistent with the results for frailty models with only parametric covariate effects, e.g., [21, 22, 23].

We used the variance component method for smoothing parameter selection in the time-varying coefficient model. Smoothing parameter selection in survival analysis can be done using cross-validation in general. O'Sullivan proposed a generalized cross-validation score for Cox models with a nonparametric covariate function [24]. Developing a cross-validation score for time-varying coefficient survival model and comparing it with the variance-component method is of future interest.

Laplace method has been used very often to approximate the integration in complicated random effects models, e.g., [12, 14, 25, 26]. In our simulation studies we showed that it performed reasonably well for survival outcomes. For sparse data, e.g., binary data, one can improve the Laplace approximation using a bias correction method [27] or inclusion of higher order terms [28]. However, the implementation of these methods is more involved. Other approaches, e.g., Monte Carlo EM algorithm, adaptive Gaussian quadrature can be also applied to this model. The implementation of these estimation methods remains a topic for our future research.

In this paper we considered the terminal event independent of the recurrent events process. This assumption may not be true. For example, Liu et al. jointly modeled the recurrent events and a terminal event using shared frailty models to account for the correlation between these two types of events [29]. We are currently extending the time-varying coefficient model for this type of joint modelling.

ACKNOWLEDGEMENTS

The project reported here was supported by the Department of Veterans Affairs, Veterans Health Administration (VHA), Office of Quality and Performance and Health Services Research and Development Service Quality Enhancement Research Initiative (RRP 09-184), an NIH grant (NIAAA RC1 AA019274), and an AHRQ grant (R01 HS020263). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs, NIH or AHRQ. We are also thankful to the associate editor and the reviewer for their constructive comments which have improved this paper substantially.

REFERENCES

1. Lin DY, Wei LJ, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. J. R. Statist. Soc. B. 2001; 62:711–730.
2. Duchateau, L.; Janssen, P. The Frailty Model. New York: Springer; 2007.
3. Cai J, Fan J, Zhou H, Zhou Y. Marginal hazard models with varying-coefficients for multivariate failure time data. Ann. Statist. 2007; 35:324–354.
4. Yu Z, Lin X. Semiparametric regression with time-varying regression coefficients for failure time data analysis. Statistica Sinica. 2009; 20:853–869. [PubMed: 20514365]
5. Sun L, Zhou X, Guo S. Marginal regression models with time-varying coefficients for recurrent event data. Statistics in Medicine. 2011 Published Online DOI: 10.1002/sim.4260.

6. Sun L, Zhu L, Sun J. Regression analysis of multivariate recurrent event data with time-varying covariate effects. Journal of Multivariate Analysis. 2009; 100:2214–2223.
7. Amorim LD, Cai JW, Zeng DL, Barreto ML. Regression splines in the time-dependent coefficient rates model for recurrent event data. Statistics in Medicine. 2005; 27(28):5890–5906. [PubMed: 18696748]
8. Chiang CT, Wang MC. Varying-coefficient model for the occurrence rate function of recurrent events. Annals of the Institute of Statistical Mathematics. 2009; 61:197–213.
9. Sun YQ, Wu HL. Semiparametric time-varying coefficients regression model for longitudinal data. Scandinavian Journal of Statistics. 2005; 32:21–47.
10. Kauermann G. Penalized spline smoothing in multivariable survival models with varying coefficients. Computational Statistics and Data Analaysis. 2005; 49:169–186.
11. Wong M, Lam KF, Lo ECM. Analysis of multilevel grouped survival data with time-varying regression coefficients. Statistics in Medicine. 2009; 30:250–259. [PubMed: 21213342]
12. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J. Am. Statist. Assoc. 1993; 88:9–25.
13. Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. J. Am. Statist. Assoc. 1992; 87:942–951.
14. Ripatti S, Palgrem J. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics. 2000; 56:1016–1022. [PubMed: 11129456]
15. Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric Regression. New York: Cambridge University Press; 2003.
16. Yu Z, Liu L. A joint model of recurrent events and a terminal event with a nonparametric covariate function. Statistics in Medicine. 2011; 30:2683–2695. [PubMed: 21751230]
17. Yao W, Barbe-Tuana FM, Llapur CJ, Jones MH, Tiller C, Kimmel R, Kisling J, Nguyen ET, Nguyen JT, Yu Z, Kaplan MH, Tepper RS. Evaluation of airway reactivity and immune characteristics as risk factors for wheezing early in life. The Journal of Allergy and Clinical Immunology. 2010; 126:483–488. [PubMed: 20816184]
18. Lichtman JH, Leifheit-Limson EC, Jones SB, Watanabe E, Bernheim SM, Phipps MS, Bhat KR, Savage SV, Goldstein LB. Predictors of hospital readmission after stroke a systematic review. Stroke. 2010; 41:2525–2533. [PubMed: 20930150]
19. Bravata D, Ordin D, Vogel B, et al. The quality of VA inpatient ischemic stroke care, FY2007: final national and medical center results of the VHA office of quality and performance (OQP) special study. 2009 2009.
20. Cook, RJ.; Lawless, JF. The Statistical Analysis of Recurrent Events. New York: Springer; 2003.
21. Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. Biometrics. 2007; 63 389397.
22. Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. Statistics in Medicine. 1995; 14:1447–1461. [PubMed: 7481183]
23. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. Statistics in Medicine. 2002; 21:3219–3233. [PubMed: 12375300]
24. O'Sullivan F. Nonparametric estimation of relative risk using splines and crossvalidation. SIAM J. Sci. Statist. Comput. 1988; 9:531–542.
25. Therneau TM, Grambsch PM, Pankratz VS. Penalized survival model and frailty. Journal of Computational and Graphical Statistics. 2003; 12(1):156–175.
26. Yu Z, Lin X, Tu W. Semiparametric frailty models for clustered failure time data. Biometrics. 2011
27. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association. 1996; 91:1007–1016.
28. Raudenbush SW, Yang M, Yosef M. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics. 2000; 9:141–157.
29. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. Biometrics. 2004; 60:747–756. [PubMed: 15339298]

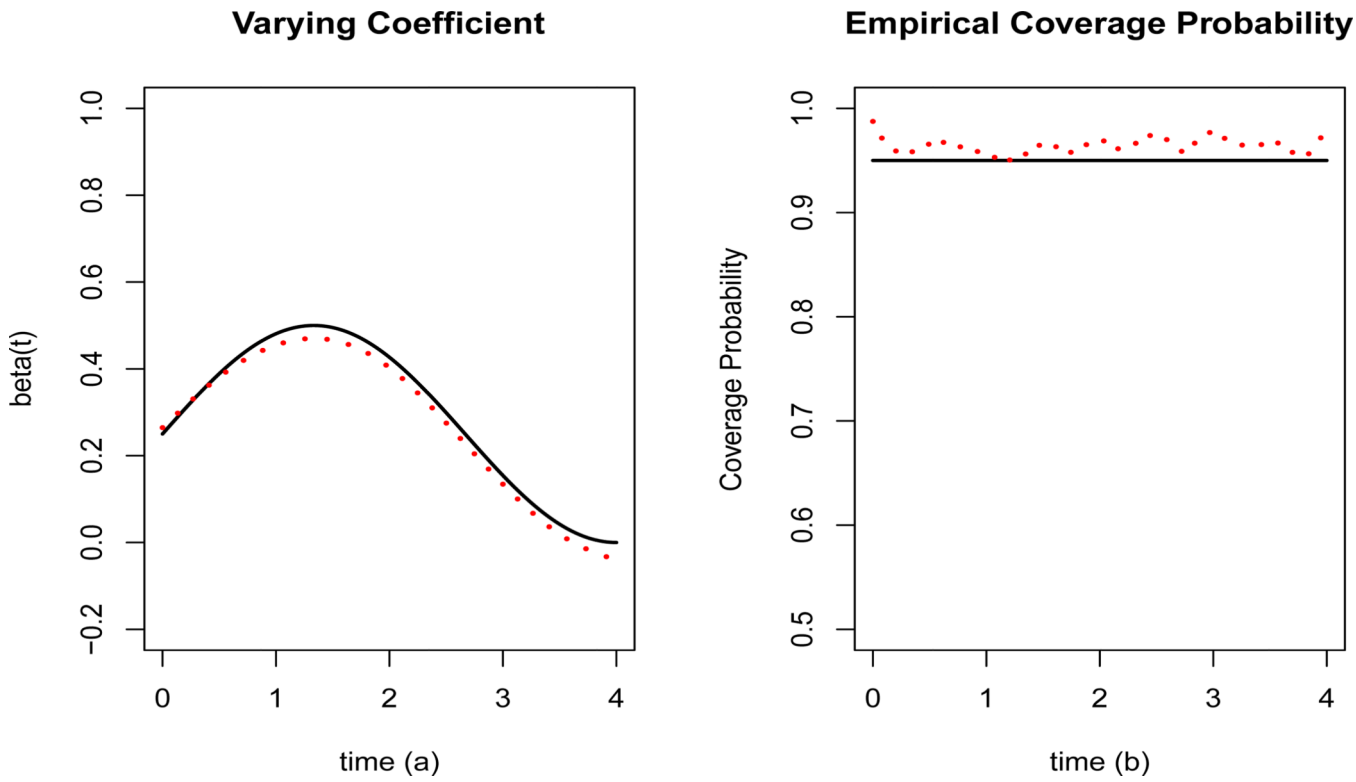

Figure 1.

Setting I: simulation results for the varying-coefficient function $\beta(t)$. (a) Penalized spline estimate of $\hat{\beta}(t)$: estimate, dotted; true $\beta(t)$, solid. (b) Point-wise empirical coverage probability of $\beta(t)$, the average coverage probability is 96.4%.

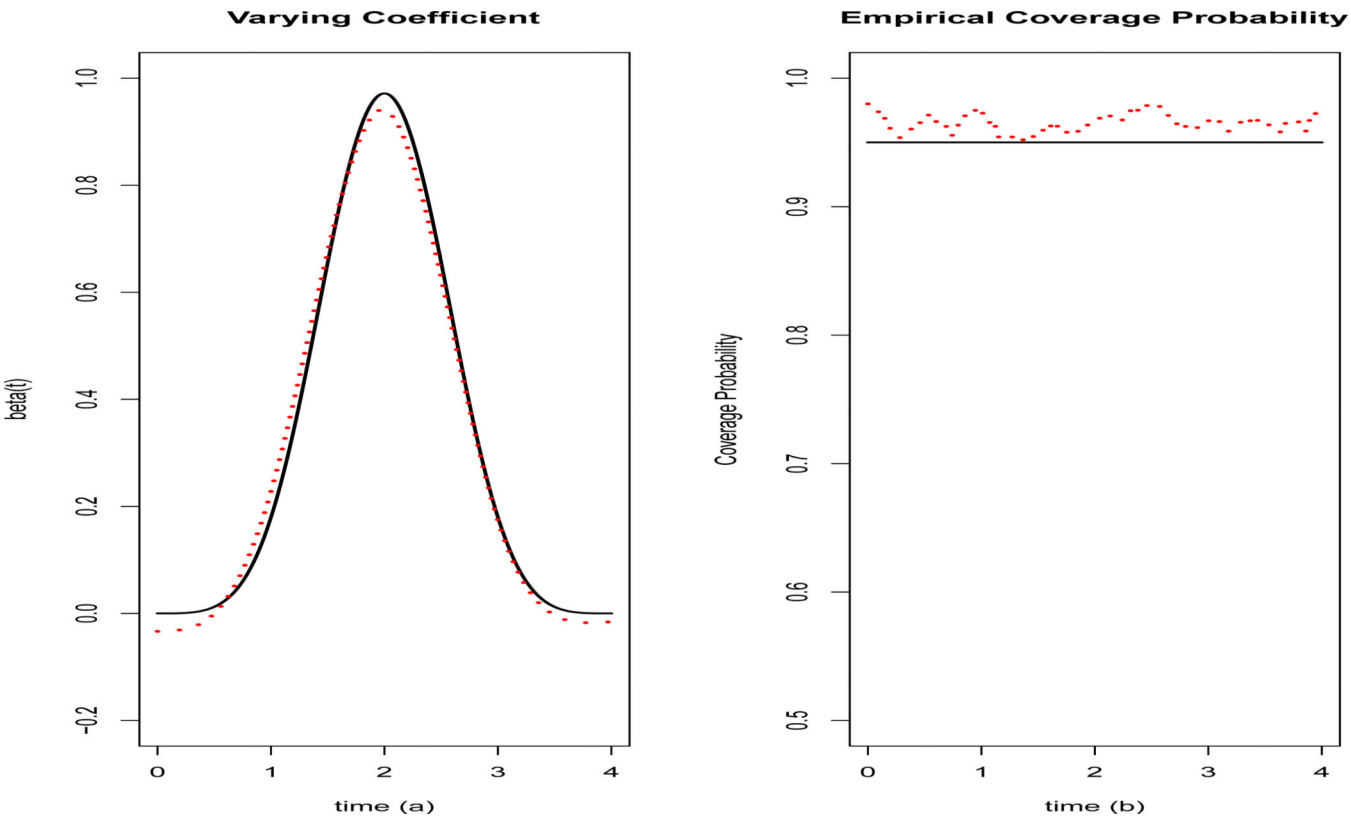

Figure 2.

Setting II: simulation results for the varying-coefficient function function $\beta(t)$. (a) Penalized spline estimate of $\beta(\hat{t})$: dotted; true $\beta(t)$, solid. (b) Point-wise Empirical coverage probability of $\beta(t)$, the average coverage probability is 96.5%.

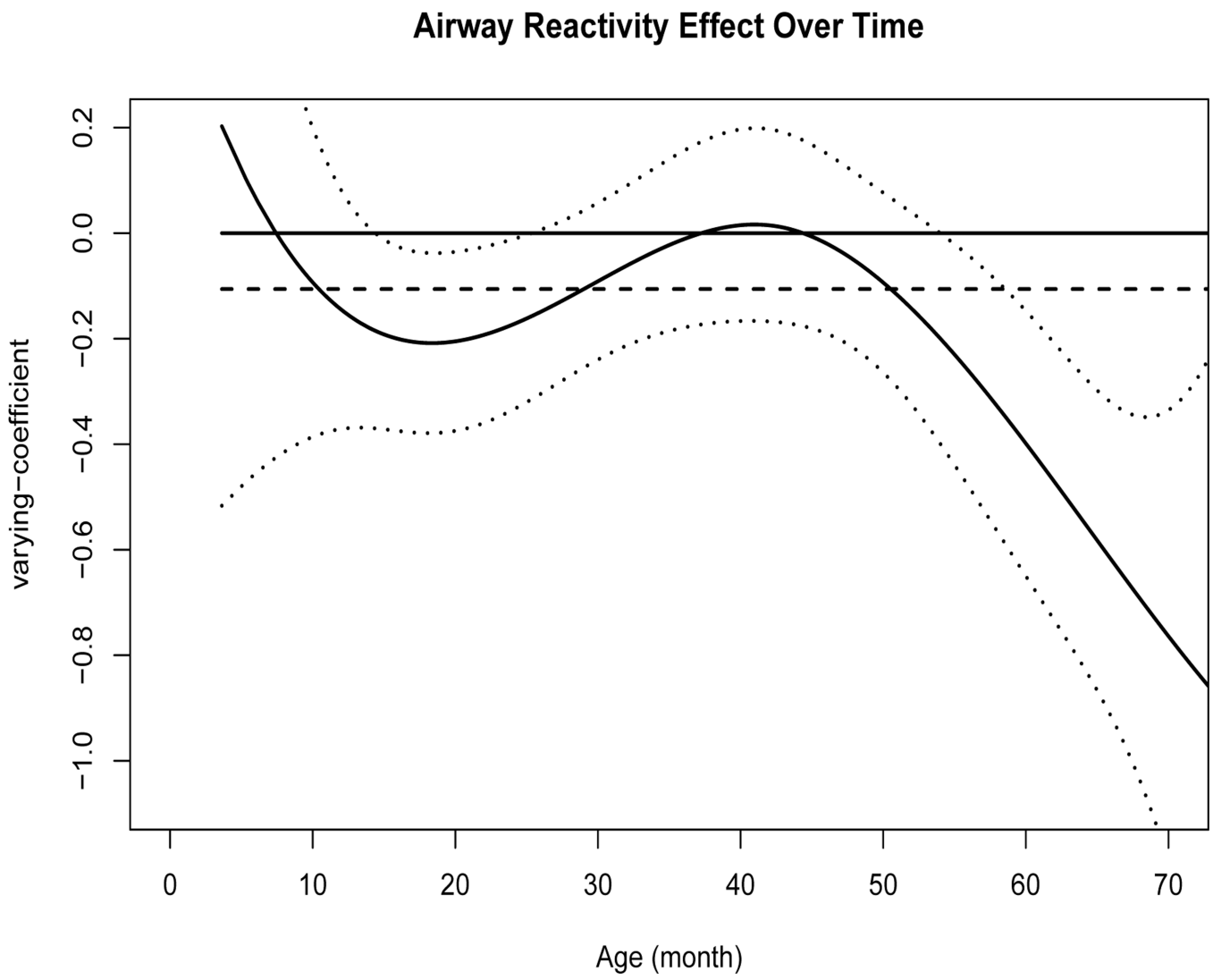

Figure 3.

Application I: Time-varying effect of log PC_{30} on recurrent wheezing: $\hat{\beta}(age)$, solid; 95% point-wise confidence interval, dotted; log PC_{30} in constant coefficient model, dashed. A horizontal line at 0 is present for better comparison.

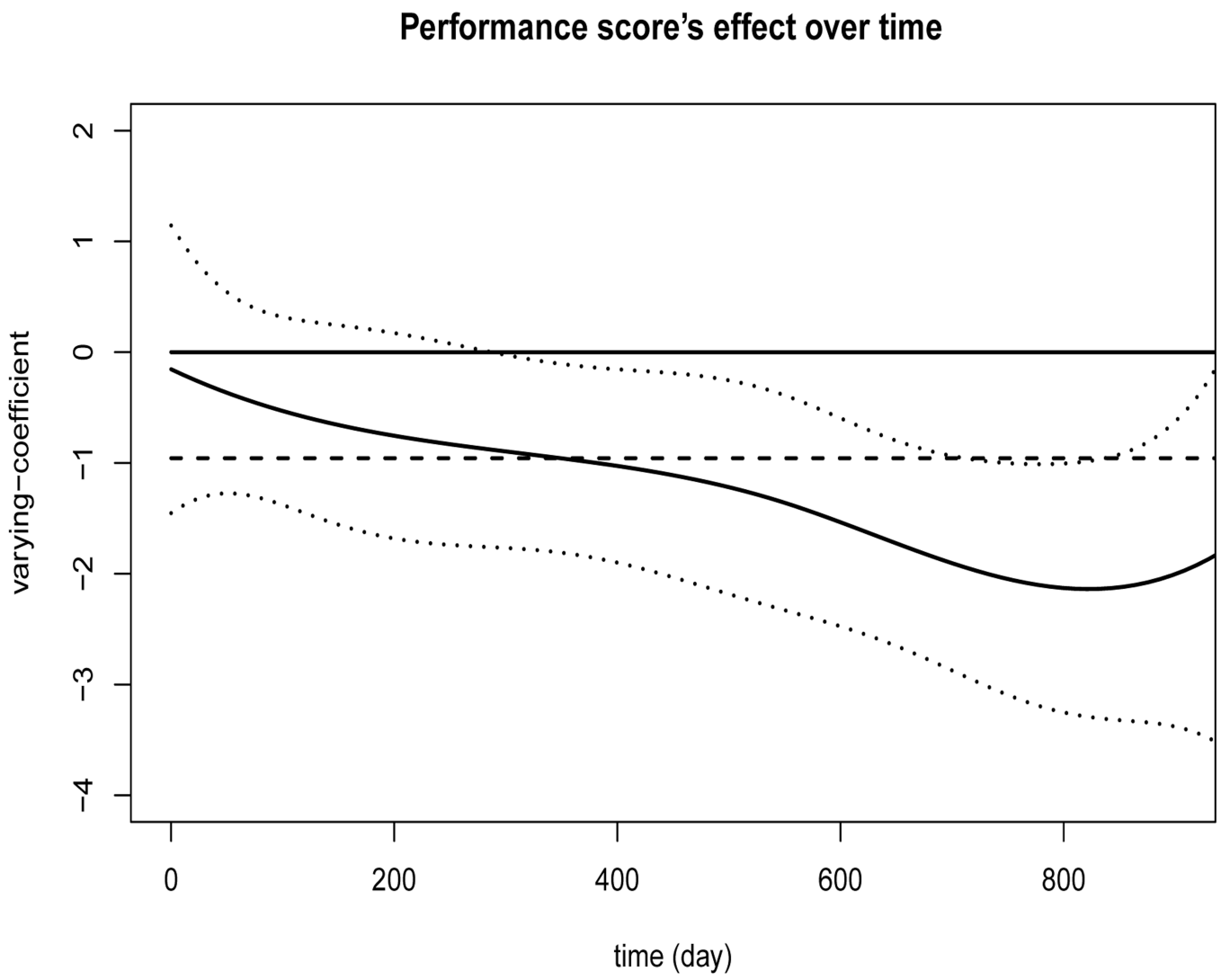

Figure 4.

Application II: Time-varying effect of the performance score on stroke readmission: $\hat{\beta}(time)$, solid; 95% point-wise confidence interval, dotted; performance score effect in constant coefficient model, dashed.

Table I

Simulation: parametric coefficient estimates

n@	m%		Param.	Mean	SE*	SEM†	CP◆
Setting I							
200	5	β_1	1	0.968	0.176	0.179	94.4%
200	5	β_2	1	0.971	0.104	0.103	94.0%
200	5	σ^2	0.5	0.465			N.A.
200	3	β_1	1	0.985	0.193	0.199	93.9%
200	3	β_2	1	0.979	0.115	0.116	93.9%
200	3	σ^2	0.5	0.456			N.A.
100	5	β_1	1	0.976	0.251	0.256	94.3%
100	5	β_2	1	0.977	0.148	0.147	94.3%
100	5	σ^2	0.5	0.455			N.A.
100	3	β_1	1	0.977	0.274	0.287	93.9%
100	3	β_2	1	0.978	0.164	0.170	93.9%
100	3	σ^2	0.5	0.436			N.A.
Setting II							
200	5	β_1	2	1.948	0.198	0.203	92.9%
200	5	β_2	2	1.947	0.130	0.129	92.9%
200	5	σ^2	0.5	0.463			N.A.
200	3	β_1	2	1.948	0.216	0.224	93.9%
200	3	β_2	2	1.963	0.146	0.151	92.9%
200	3	σ^2	0.5	0.460			N.A.
100	5	β_1	2	1.947	0.283	0.298	93.1%
100	5	β_2	2	1.954	0.186	0.184	92.8%
100	5	σ^2	0.5	0.452			N.A.

n@	m%		Param.	Mean	SE*	SEM†	CP◇
100	3	β_1	2	1.952	0.306	0.316	92.8%
100	3	β_2	2	1.960	0.208	0.218	93.3%
100	3	σ^2	0.5	0.436			N.A.

@sample size;

%number of events per subject;

*Empirical standard error;

†Mean of standard error;

◇Coverage probability.

Table II

Parametric Coefficient Estimates of Childhood Wheezing

Full model				
Risk factors	**Est.**	**SE**	**Hazards ratio**	**p-value**
Pregnancy Smoking	0.138	0.030	1.820	<0.0001
Gender (Male)	0.303	0.099	1.35	0.002
σ	0.12			

Table III

Parametric Coefficient Estimates of Stroke Readmission Model

Risk factors	Est.	SE	Hazards ratio	p-value
Charlson's Score	0.189	0.030	1.208	<0.0001
Smoker	0.103	0.127	1.108	0.421
σ	0.320			